Supplementary Information

The integration of motion and disparity cues to depth in dorsal visual cortex

Hiroshi Ban, Tim J Preston, Alan Meeson and Andrew E Welchman

Overview

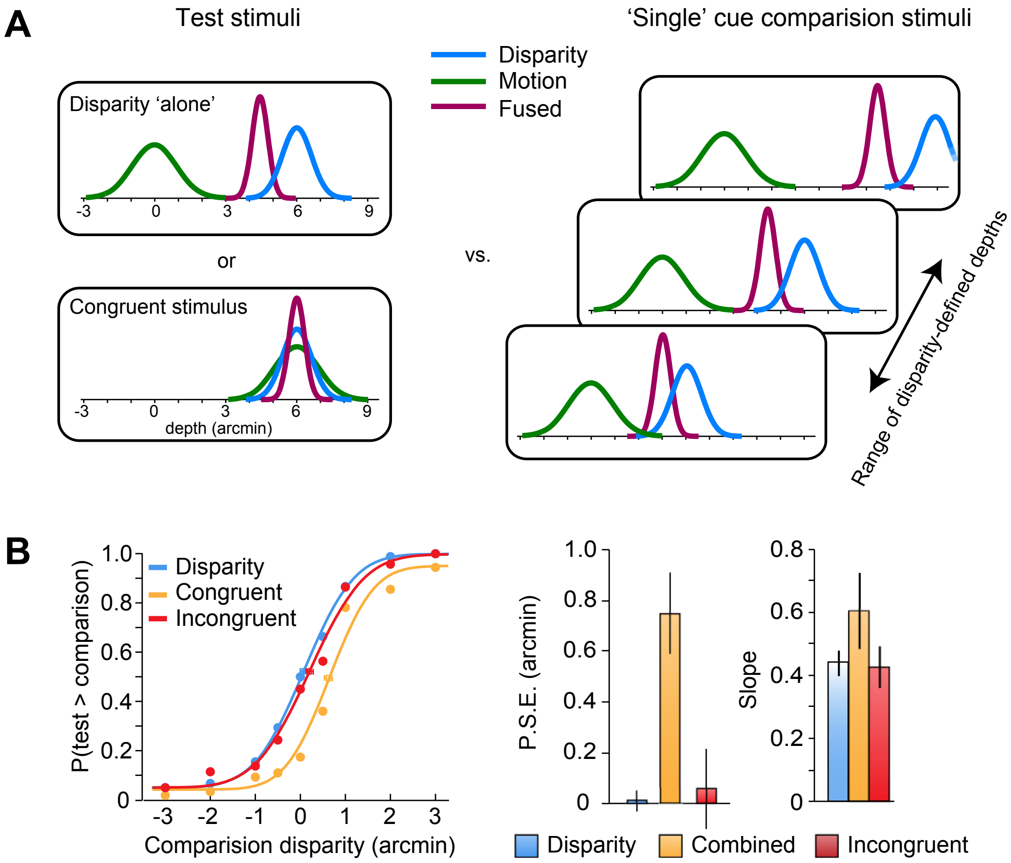

Figure S1 | Behavioral results from additional psychophysical tests

(A) To test whether information from both cues was combined behaviorally, we used a psychophysical procedure in which two stimuli were presented sequentially, and participants had to decide which had the greater depth. The visual stimuli consisted random dot patterns (Main **Fig. 2b**) that depicted depth structure defined by: (1) a difference in binocular disparity, (2) the congruent- and (3) the incongruent combination of disparity and motion. The stimulus parameters (size, dot density, motion speed etc.) matched the stimuli used for fMRI experiments. On each trial, a 'test' and a 'comparison' stimulus were shown sequentially for 1 s each in a random order, with a 1 s interstimulus interval. The relative depth in the 'comparison' stimulus was specified by disparity, while the depth in the 'test' stimulus could be specified by disparity, or disparity and motion in combination (only the congruent case is illustrated, but incongruent stimuli were also shown). By contrasting a given test stimulus against a range of comparison stimuli, we obtained psychometric functions. These expressed the perceptual likelihood that depth in the test stimulus exceeded depth in the comparison stimulus– where depth is expressed in terms of the perception of depth in the disparity 'alone' (i.e. conflicting) comparison stimuli.

(B) We found that participants (N = 8) reported greater depth when disparity and motion congruently indicated depth differences. This is shown by the rightwards shift in the orange psychometric function relative to the blue, baseline curve for the group data (horizontal boxplots on the functions show the error associated with the point of subjective equality (P.S.E.) of the group data). This shift indicated that cues were integrated to informed depth percepts. We quantified this across subjects using the P.S.E. (bar graphs show between-subjects mean based on fits to individual subjects' data; error bars depict s.e.m.). The P.S.E. was reliably greater when motion and disparity signaled the same depth structure ($F_{1,7} = 21.14$, *P* = 0.002). Note that this does not suggest that cue fusion seeks to increase the magnitude of depth estimates (i.e. adding more and more cues does not lead to greater and greater depth)– rather bear in mind that the comparison stimuli contain cue conflicts, and thus the perceptual interpretation of the comparison stimuli is biased (towards zero) away from the disparity-specified depth (e.g. 6 arcmin). (To appreciate this visually, compare the illustrations of the test stimuli in part A, in which disparity—blue curve—specifies the same depth in the two cases, but the perceptual estimate—purple curve—is greater for the congruent stimulus.) Thus when contrasting conflicting single cue stimuli against the congruent disparity and motion condition, depth for the congruent stimulus exceeds the value specified by the 'single' cue disparity stimulus. In addition to changing the P.S.E., the slopes of the psychometric functions were steeper for congruent depth cues ($F_{2,14}$ = 3.26, *P* = 0.035). This is shown by the between-subjects mean slope bar graphs (error bars show s.e.m.), and is expected on the basis that integration improves the reliability of depth estimates.

This influence on the P.S.E. and sensitivity was specific to congruent combinations of cues (compare orange and red bar graphs). Specifically, incongruent PSEs differed from congruent (*P* = 0.009) but not disparity alone (*P* = 0.401) conditions; and the slope of the psychometric function was lower for incongruent cues relative to congruent cues (*P* = 0.041), but no different compared to the disparity alone condition (*P* = 0.398).

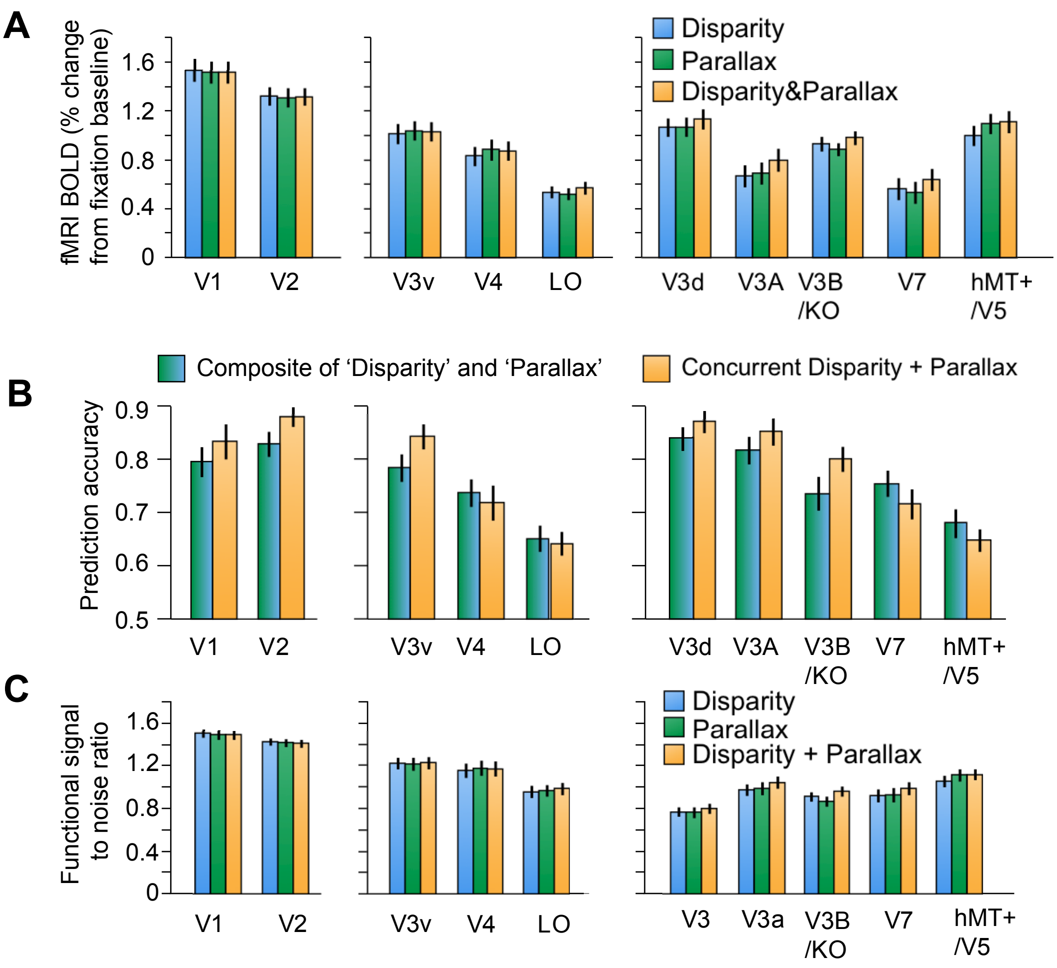

Figure S2 | Percent signal change and composite data

(A) Percent signal change from fixation baseline across conditions. We computed the percent fMRI signal change (%) for each condition relative to the fixation baseline response for each participant and Region of Interest (ROI). This was calculated from the mean response of the 250 voxels used for classification in each ROI. Error bars show between-subject s.e.m. (N = 20).

(B) Prediction accuracies for composite vs. congruent conditions. To evaluate the idea that presenting two signals concurrently reduced measurement noise, we generated a composite dataset that averaged together responses from 'single' cue conditions. For each voxel, we averaged the fMRI response evoked under the 'disparity' and 'motion' conditions. We then ran multivoxel pattern analysis (MVPA) using this composite data and compared decoding performance with that supported by the fMRI signals evoked by the congruent disparity and motion condition. If the two depth cues are represented independently in a cortical region, we might expect no difference in the decoding performance between the composite and congruent conditions. In contrast, if depth information from two different cues is integrated optimally, noise will decrease, with the result that performance for the congruent case will improve relative to the composite data. We found a significant interaction between ROI and condition ($F_{4,69}$ = 2.491, P = 0.04), with post-hoc testing revealing that decoding performance in V3B/KO for the congruent case was significantly above that obtained for the composite data. This suggests that simple measurement noise cannot account for our findings, and supports an interpretation based on cue integration. Differences in other areas were not significant based on correction for multiple comparisons.

(C) Functional signal to noise ratios (fSNR). Functional signal to noise values for the 250 voxels used for classification per ROI. fSNR was defined as:

$$SNR = \frac{x_{Stimulation} - x_{Fixation}}{STD(x_{Stimulation\&Fixation})}$$

where $x_{Stimulation} - x_{Fixation}$ represents the difference between the mean response to the stimuli and response to fixation, and $STD(x_{Stimulation\&Fixation})$ the standard deviation across all stimulus conditions and fixation. The pattern of SNR does not correlate with classification accuracy.

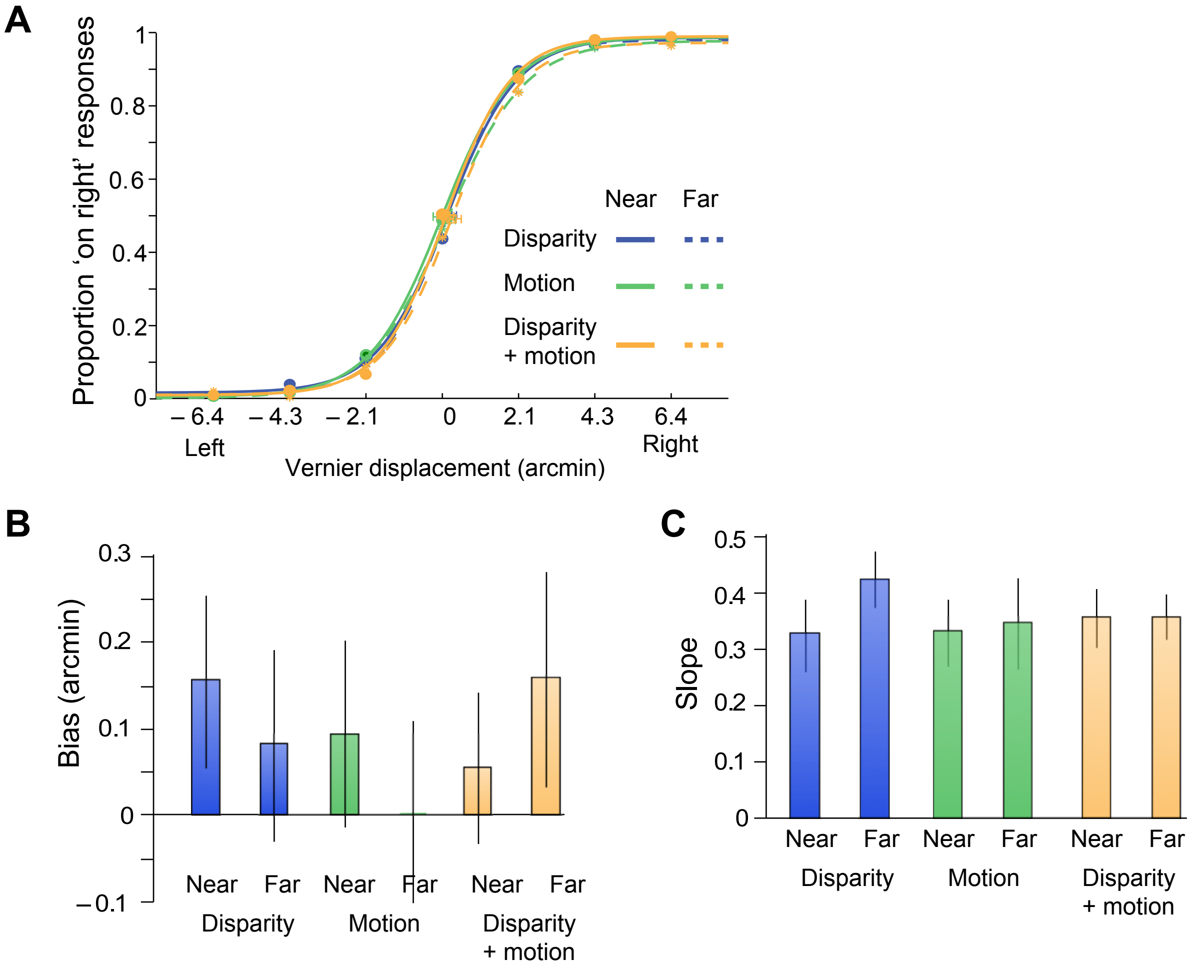

Figure S3 | Subjective assessment of eye vergence

(A) The behavioral task during scanning consisted of a subjective assessment of eye vergence task (Popple *et al.* 1998). A small vernier target was briefly (250 ms) flashed to one eye, and participants judged its location relative to the upper vertical nonius line (presented to the other eye). We fit the proportion of "*target is to the right of the upper nonius line*" responses as a function of the target's displacement. These data were fit separately for near and far positions in each experimental condition.

(B) The mean bias term (P.S.E.) for the fits to the psychophysical data under each condition and for near and far positions. Error bars show the between-subjects s.e.m. (N = 20). A repeated-measures ANOVA showed no main effect of condition ($F_{2,38} < 1$, P = 0.45) or depth position ($F_{1,19} < 1$, P = 0.73) and no interaction ($F_{2,38}$ = 1.38, P = 0.26). These data suggest that participants were able to maintain vergence at the fixation point well as mean bias is very close to zero.

(C) The mean slope term for fits to the psychophysical data under each condition for near and far positions. Error bars show the between-subjects s.e.m. (N = 20). A repeated-measures ANOVA showed no main effect of condition ($F_{1.4,26.9} < 1$, P = 0.59) or depth position ($F_{1,19}$ = 1.98, P = 0.18) and no interaction ($F_{2,38} < 1$, P = 0.40).

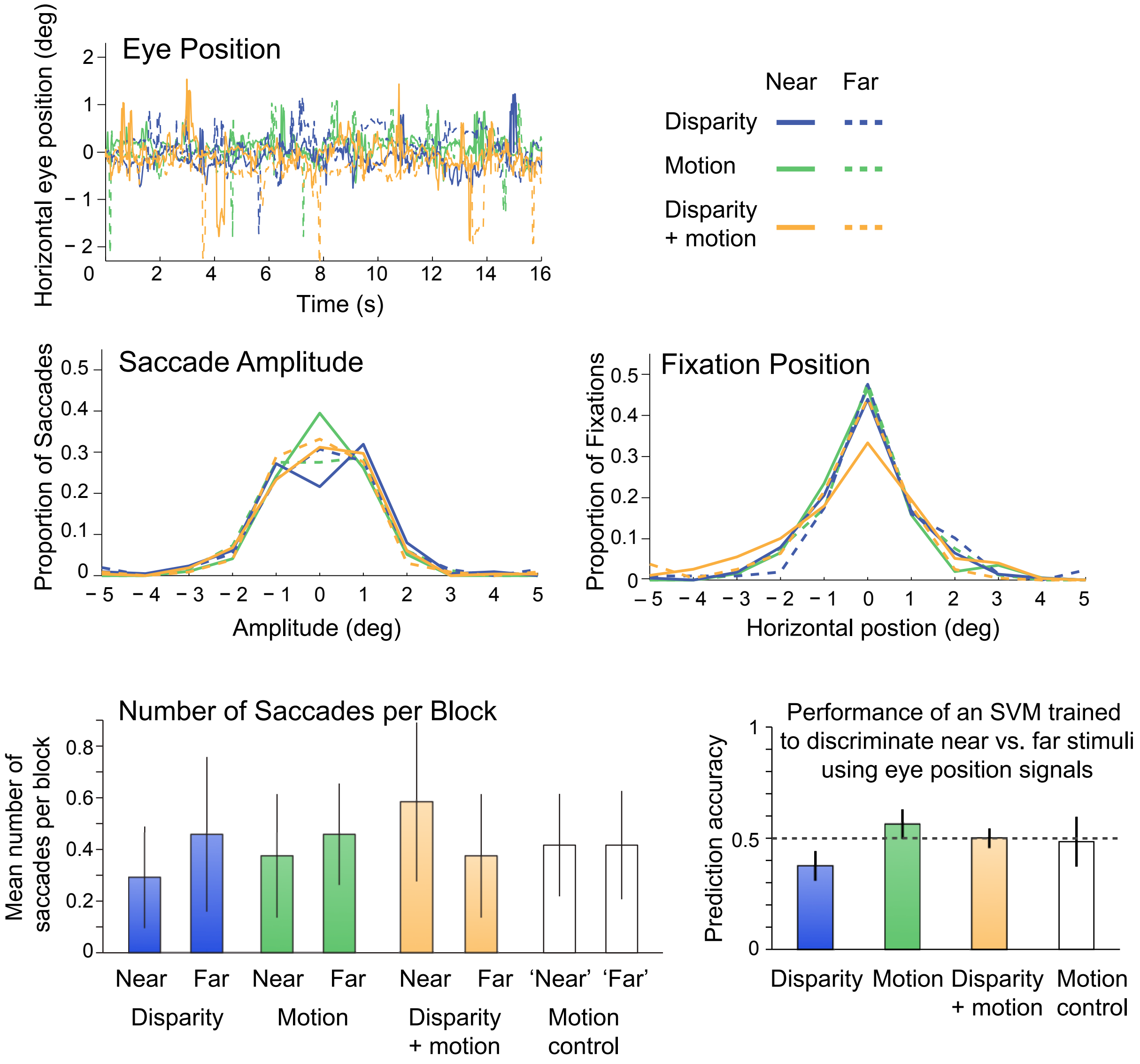

Figure S4 | Eye movement recordings

We recorded horizontal eye-movements from four subjects at a high resolution (stated accuracy < 0.25 degrees visual angle) using a monocular limbus eye-tracker (CRS Ltd, Rochester, UK) that was placed under the spectral interference filters inside the bore of the magnet. This eye tracker is the only system compatible with the dual-projector system we used for binocular presentation, as video-based systems cannot reliably track the eye through the spectral comb filters. No significant differences were observed across conditions and experiments in the mean eye position ($F_{2,6} < 1$, $P = 0.86$), number of saccades ($F_{2,6} < 1$, $P = 0.52$) and saccade amplitude ($F_{2,6} = 3.44$, $P = 0.10$).

Traces of mean eye position aligned to the start of each trial showed only small deviations from fixation and no systematic differences between conditions. Despite technical limitations (that is, it was only possible to measure horizontal position in one eye), these results suggest that observers could maintain fixation throughout each run. Using trial-by-trial eye movement traces, we trained an SVM to associate patterns of eye movement with the 'near' or 'far' position of the viewed stimulus. We assessed the prediction performance of the SVM based on these eye position signals, and found that performance did not differ significantly from chance (0.5), making it unlikely differences in eye position account for our results.

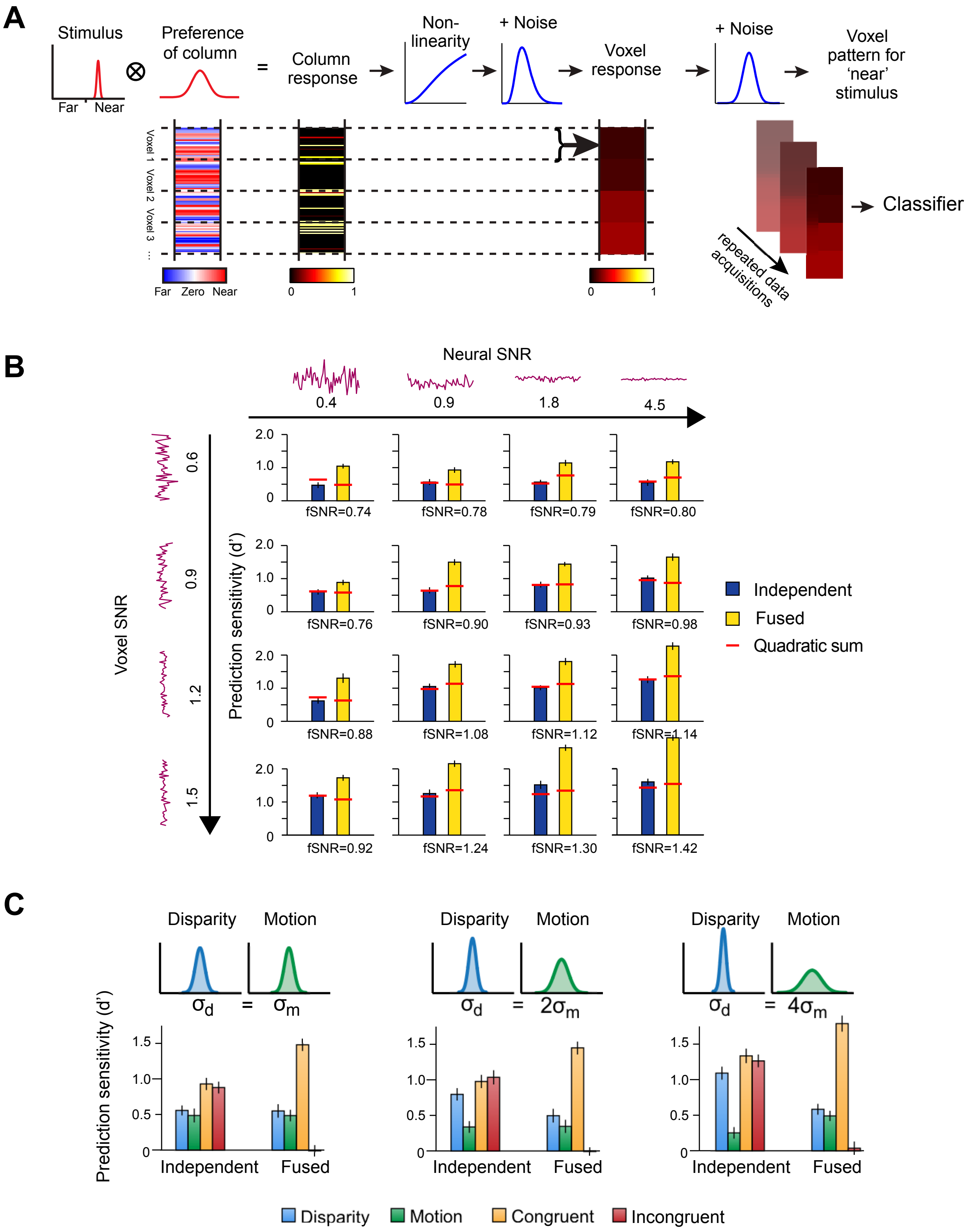

Figure S5 | Simulation results – effects of SNR and cue reliability

(A) To confirm our experimental logic, we performed simulations of voxel responses that were decoded by a Support Vector Machine (SVM). We simulated a population of 'depth columns', each of which had a mean depth preference and a fixed tuning width. Columns had a spatial sawtooth structure whose phase progression was randomly perturbed to create jittered maps following Kamitani and Tong (2005). Columns could respond to disparity, motion, or combined signals and thereby reduce their tuning variability following maximum likelihood estimation. The default column tuning profiles were assumed to be Gaussian with σ = 12 arcmin. The stimulus was represented by a Gaussian (σ = 0.2 arcmin). We simulated the response of these individual columns when presented with depth structures (± 6 arcmin) defined by disparity, motion and these signals concurrently (that is, convolution of the probability density functions associated with the stimulus and the tuning profile of the column). These column responses were subject to an expansive–compressive saturating non-linearity in their response following the approach of Boynton, Demb, Glover & Heeger (1999). Column responses were subject to 'neural noise' in the form of an added random value sampled from a Gamma distribution (k = 9, θ = 0.05 by default).

To calculate voxel responses, we averaged the responses of individual columns that were sampled by a coarser scale voxel grid (1.5 mm length, matching the fMRI scans). We assumed that each voxel sampled approximately half a spatial period of the underlying depth map (one columnar cycle was set to 3.0 mm) based on scaling for human cortex relative to disparity representations in macaque MT (DeAngelis & Newsome, 1999). These aggregated column responses were then subjected to 'voxel noise' in the form of random values sampled from a normal distribution (σ = 0.5 by default). The added voxel noise had two components – one was a (quasi) unique noise value; the other was a correlated component common to each voxel at a given time point (i.e. to reflect global fluctuations in fMRI noise across time). The ratio of the correlated and random noise was fixed to 1:9 for all the simulations.

Neural and fMRI noise parameters used in the simulations were carefully selected so that the signal-to-noise ratios (SNRs) of the final voxel responses matched the empirical fMRI data. Neural SNR, voxel SNR and functional SNR (fSNR) ratio were separately defined following the formula described below.

$$neural\ SNR = \frac{mean(A_{neural\ response})}{STD(A_{neural\ response\ \&\ noise})}$$

$$voxel\ SNR = \frac{mean(A_{voxel\ response} - A_{neural\ noise})}{STD(A_{voxel\ response\ \&\ voxel\ noise})}$$

$$fSNR = \frac{mean(A_{voxel\ response})}{STD(A_{voxel\ response\ \&\ voxel\ noise})}$$

A indicates response amplitude of each source. Simulated data with an fSNR of 0.93 matched with the SNR estimated from the fMRI data in V3B/KO. This value was used for simulations, unless specified otherwise.

We simulated responses of voxels that contained (i) populations of columns that fused information from disparity and motion, and (ii) voxel responses that contained co-located but independent populations (i.e. voxel responses were driven by populations that responded to only disparity or only motion). We organized maps such that there was a similar topographical structure for the two depth cues. The process of sampling from these maps entailed that the distribution of column selectivities differed for the two cues for a given voxel.

We simulated 250 voxels with 8 runs of 24 patterns for both near and far presentations for each condition. These numbers matched fMRI data acquisition. The simulated patterns were then sent to a linear SVM classifier using leave-one-out cross validation procedures. The SVM classifications were repeated 20 times, simulating 20 participants in the main fMRI experiment, and then averaged.

(B) We manipulated SNRs of simulated fMRI responses to evaluate how the noise levels affected our results. Specifically, we used a range of neural (0.4, 0.9, 1.8, 4.5) and voxel (0.6, 0.9, 1.2, 1.5) SNRs, that together defined the functional SNR (ranging from 0.74 to 1.42). The bar graphs show performance for Fused and Independent populations for the congruent-cue stimulus. Performance for 'single' cue conditions was also calculated (separately for the two different populations) to generate the quadratic summation prediction (red line on the graphs). The simulations reveal that fused populations always exceed the quadratic summation prediction. Independent populations may slightly surpass the quadratic summation prediction in cases of low noise. This is possibly due to partial correlations emerging between voxels when voxel responses are dominated by column responses (i.e. low levels of independent late noise); that is, the same voxel set is used for each condition and these voxels are assumed to contain columns selective for both cues, where cue maps have a correlated topography. In consequence, the summation test can establish a minimum bound for fusion, but does not preclude independent populations. For the remainder of the simulations we used a neural SNR of 1.8, and a voxel SNR of 0.90, giving an fSNR of 0.93 that matched empirical data.

(C) We tested how classification accuracies changed when the disparity and motion cue reliabilities differed. We changed the standard deviation of column response to motion so that it was 1.0, 2.0, and 4.0 times larger than of disparity, while the variance of an integrated response was held constant. While this manipulation affects decoding performance considerably for the independent population model, performance of a fused mechanism is little affected.

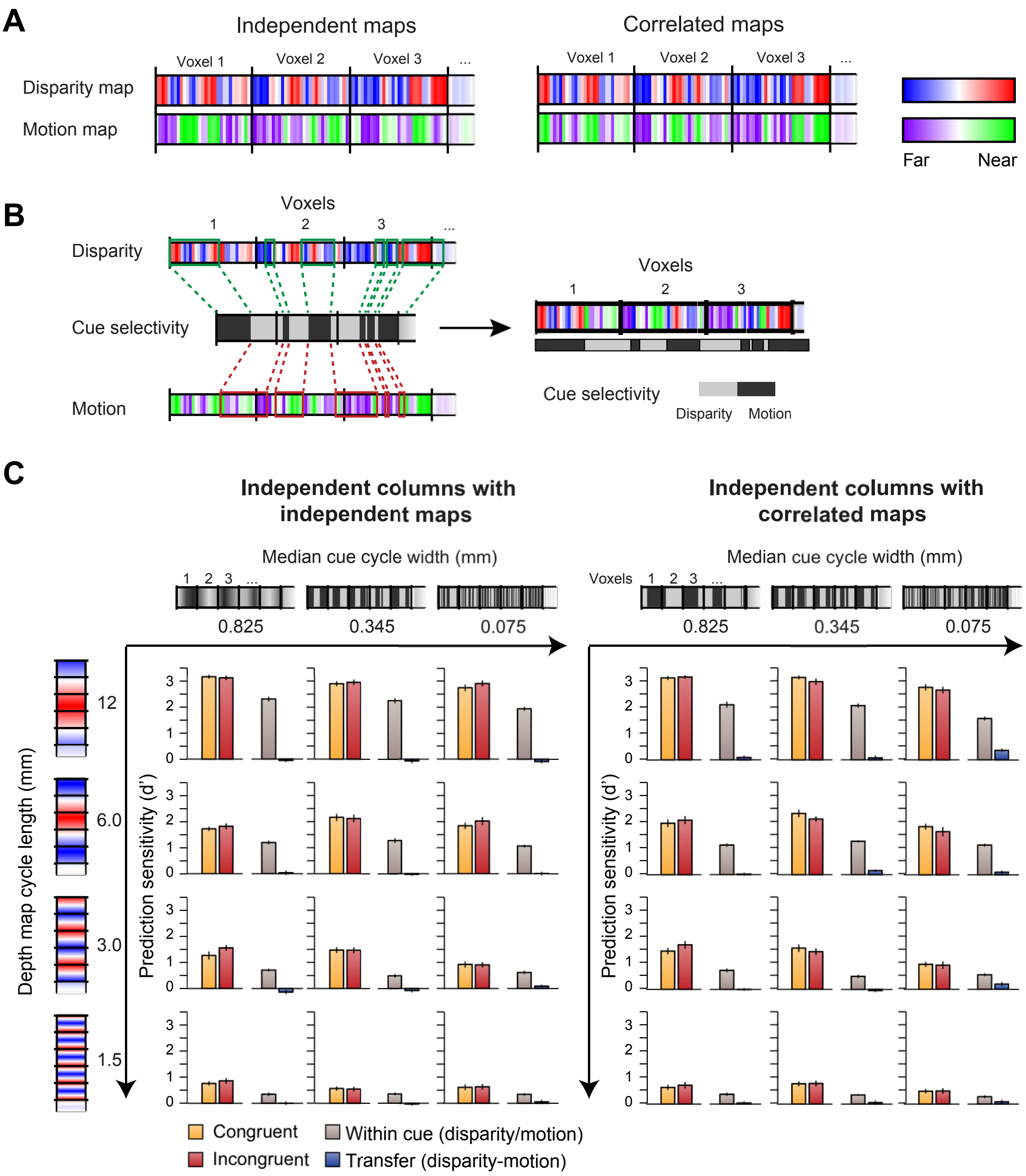

Figure S6 | Simulation results – effects of spatial organization

(A) We tested how the relative organization of disparity and motion maps might interact within a voxel, by simulating two extreme scenarios: complete correlation and complete independence (correlation between maps = 0). Color-coded bars show the peak preference of the depth columns.

(B) We tested how maps from disparity and motion are sampled by individual voxels. We simulated different scales of sampling from each map to produce the pattern of column responses that are aggregated to produce the voxel response. Sampling widths were defined by a sawtooth structure whose phase progression was randomly perturbed– this is illustrated by the 'cue selectivity' color bar.

(C) Results of simulations for independent and correlated neuronal populations for the case where disparity and motion are independently represented. For the two scenarios, we present a grid of results where we varied systematically the spatial period of depth maps (global *y*-axis) and the spatial period of the cue selectivity (global *x*-axis). These manipulations are shown by schematics next to the axes; note the depth maps cycles are illustrated as not jittered for clarity of presentation; however, the maps were jittered for the simulations. The results in each cell of the grid show decoding performance for the congruent and incongruent conditions, and performance in the transfer test. It is readily appreciated that for both correlated and independent scenarios, independent populations do not give rise to reliable differences between congruent and incongruent cues, or produce reliable between-cue transfer effects, unlike fused representations (Main **Fig. 6**).

We varied the spatial scale of the depth maps around the value of 3 mm generally used for the simulations. In particular, we considered a higher spatial frequency (1.5 mm, approximately equivalent to disparity maps in the macaque) and two lower spatial frequencies (6, 12 mm). As the scale of the maps increases, voxels obtain a more homogenous sample of columnar selectivities. However, even with very large-scale representations (i.e. 12 mm maps that would take up about half the ROI; mean maximum diameter of V3B/KO in six participants is 27.3 ± 2.3 mm), we do not observe reliable congruent vs. incongruent differences or transfer. We also changed the width of the maps of cue selectivity from the default value of 0.825 mm (a conservative estimate based on data for ocular dominance and orientation in V1 Yacoub et al, 2008; Kriegeskorte et al, 2010) to higher scale representations (0.345, 0.075 mm). As cue selectivity scale decreases, a voxel's sample of columnar responses becomes more similar. Even with unrealistically high interdigitation of individual cues, independent populations do not give rise to patterns of performance associated with fusion.

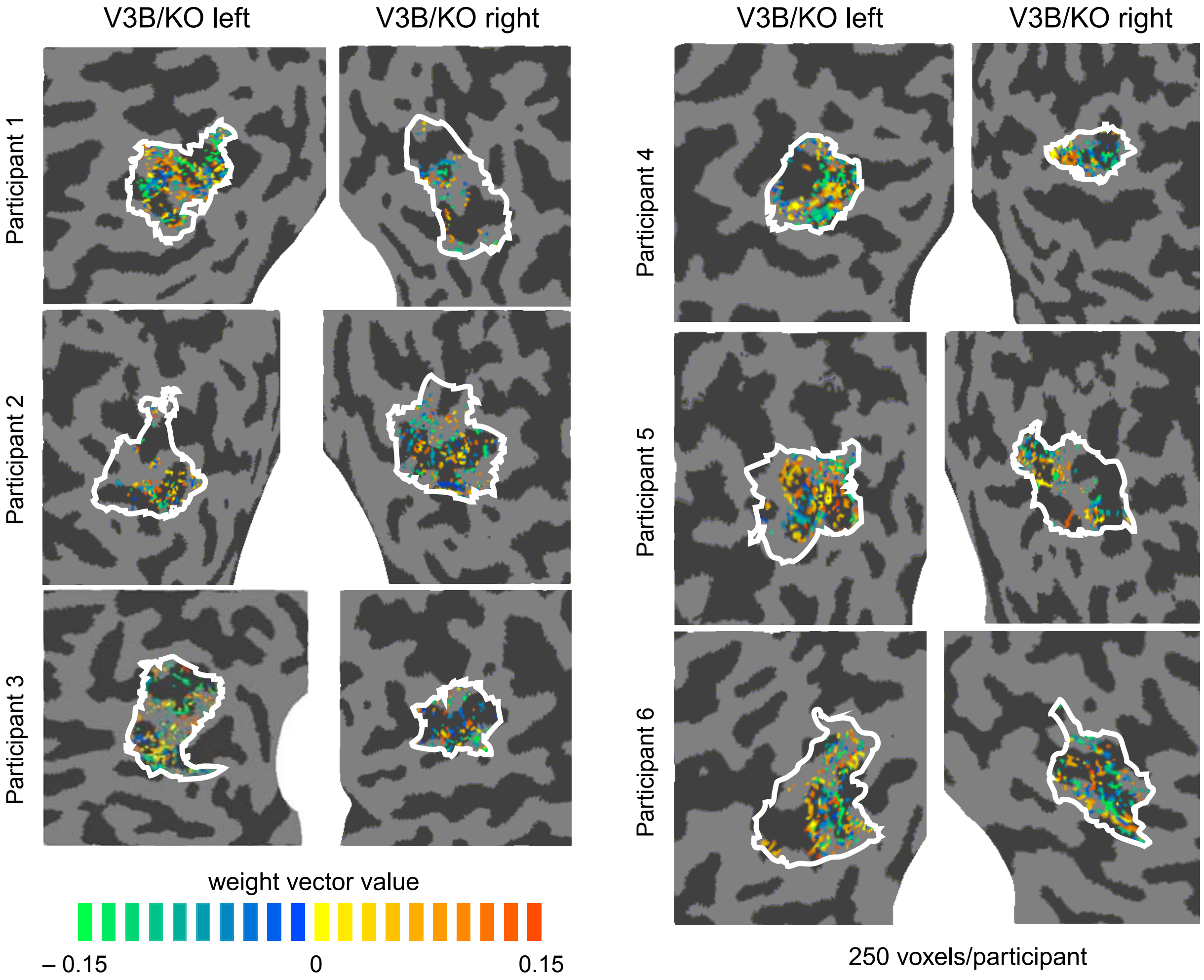

Figure S7 | The organization of weighted voxels within V3B/KO

The V3B/KO region of interest (white outline) showing the spatial location of voxels used by the SVM learning algorithm (color coded by the weight given to the voxel). Portions of the dorsal and lateral visual cortex are shown for 6 participants on flattened representations of each hemisphere. These maps were created using the same approach as the global maps illustrated in Main **Fig. 3**. It is evident that there is no consistent spatial clustering in the voxels used by the SVM, as might be expected if there were two distinct representations (i.e. V3B *vs*. KO) within the region of interest we denote as V3B/KO.

References

Boynton GM, Demb JB, Glover GH & Heeger DJ (1999) Neuronal basis of contrast discrimination *Vision Res* **39**, 257-69

DeAngelis GC & Newsome WT (1999) Organization of disparity-selective neurons in macaque area MT *J Neurosci* **19**, 1398-415

Kamitani Y & Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci* **8**, 679-85

Kriegeskorte N, Cusack R & Bandettini P (2010) How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? *NeuroImage* **49**, 1965-76

Popple AV, Smallman HS & Findlay JM (1998) The area of spatial integration for initial horizontal disparity vergence. *Vision Research* **38**, 319-326.

Yacoub E, Harel N & Ugurbil K (2008) High-field fMRI unveils orientation columns in humans. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 10607–12